Published online 16 November 2006 Nucleic Acids Research, 2007, Vol. 35, Database issue D193–D197
doi:10.1093/nar/gkl929

The Universal Protein Resource (UniProt)

The UniProt Consortium[1,2,3]

[1]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007, USA, [2]The EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK and [3]Swiss Institute of Bioinformatics, Centre Medical Universitaire 1 rue Michel Servet, 1211 Geneva 4, Switzerland

Received September 21, 2006; Accepted October 12, 2006

ABSTRACT

The ability to store and interconnect all available information on proteins is crucial to modern biological research. Accordingly, the Universal Protein Resource (UniProt) plays an increasingly important role by providing a stable, comprehensive, freely accessible central resource on protein sequences and functional annotation. UniProt is produced by the UniProt Consortium, formed in 2002 by the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). The core activities include manual curation of protein sequences assisted by computational analysis, sequence archiving, development of a user-friendly UniProt web site and the provision of additional value-added information through cross-references to other databases. UniProt is comprised of three major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase and the UniProt Reference Clusters. An additional component consisting of metagenomic and environmental sequences has recently been added to UniProt to ensure availability of such sequences in a timely fashion. UniProt is updated and distributed on a bi-weekly basis and can be accessed online for searches or download at http://www.uniprot.org.

INTRODUCTION

High-throughput genome sequencing is producing a rapid and accelerating accumulation of predicted protein sequences for a large number of organisms. At the same time, protein functions are being analyzed using a wide range of approaches, ranging from traditional small-scale experiments to large-scale methods such as gene expression profiling, protein-protein interactions and structural genomics as well as *in silico* prediction of protein functions. To accommodate these data, various individual resources are available to the research community. However, there is a widely recognized need for a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, integrating and standardizing data from these various sources.

UniProt is the central resource for storing and interconnecting information from large and disparate sources and the most comprehensive catalog of protein sequence and functional annotation. It has three components optimized for

To whom correspondence should be addressed to Rolf Apweiler. Tel: +44 1223 494435; Fax: +44 1223 494468; Email: rolf.apweiler@ebi.ac.uk

The UniProt Consortium: Amos Bairoch, Lydie Bougueleret, Severine Altairac, Valeria Amendolia, Andrea Auchincloss, Ghislaine Argoud Puy, Kristian Axelsen, Delphine Baratin, Marie-Claude Blatter, Brigitte Boeckmann, Laurent Bollondi, Emmanuel Boutet, Silvia Braconi Quintaje, Lionel Breuza, Alan Bridge, Edouard deCastro, Danielle Coral, Elisabeth Coudert, Isabelle Cusin, Pavel Dobrokhotov, Dolnide Dornevil, Severine Duvaud, Anne Estreicher, Livia Famiglietti, Marc Feuermann, Sebastian Gehant, Nathalie Farriol-Mathis, Serenella Ferro, Elisabeth Gasteiger, Alain Gateau, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nicolas Hulo, Vassilios Ioannidis, Ivan Ivanyi, Janet James, Eric Jain, Silvia Jimenez, Florence Jungo, Vivien Junker, Guillaume Keller, Corinne Lachaize, Lydie Lane-Guermonprez, Petra Langendijk-Genevaux, Vicente Lara, Philippe Lemercier, Virginie Le Saux, Damien Lieberherr, Tania de Oliveira Lima, Veronique Mangold, Xavier Martin, Karine Michoud, Madelaine Moinat, Cristiano Moreira, Anne Morgat, Marisa Nicolas, Shoko Ohji, Salvo Paesano, Ivo Pedruzzi, David Perret, Isabelle Phan, Sandrine Pilbout, Violaine Pillet, Sylvain Poux, Nicole Redaschi, Sorogini Reynaud, Catherine Rivoire, Bernd Roechert, Claudia Sapsezian, Michel Schneider, Christian Sigrist, Mauricio da Silva, Karin Sonesson, Andre Stutz, Shyamala Sundaram, Michael Tognolli, Laure Verbregue, Anne-Lise Veuthey, Claudia Vitorello and Lina Yip at the Swiss Institute of Bioinformatics (SIB) and the Medical Biochemistry Department of the University of Geneva; Rolf Apweiler, Yasmin Alam-Faruque, Daniel Barrell, Lawrence Bower, Paul Browne, Wei Mun Chan, Louise Daugherty, Emilio Salazar Donate, Ruth Eberhardt, Alexander Fedotov, Rebecca Foulger, Gill Fraser, Gabriella Frigerio, John Garavelli, Renato Golin, Alan Horne, Julius Jacobsen, Michael Kleen, Paul Kersey, Ernst Kretschmann, Kati Laiho, Rasko Leinonen, Duncan Legge, Michele Magrane, Maria Jesus Martin, Patricia Monteiro, Claire O'Donovan, Sandra Orchard, John O'Rourke, Samuel Patient, Manuela Pruess, Andrey Sitnov, Nataliya Sklyar, Eleanor Whitfield, Daniela Wieser, Quan Lin, Mark Rynbeek, Giuseppe di Martino, Mike Donnelly and Pieter van Rensburg at the European Bioinformatics Institute (EBI). Cathy Wu, Cecilia Arighi, Leslie Arminski, Winona Barker, Yongxing Chen, Sehee Chung, Christina Fang, Vincent Hermoso, Zhang-Zhi Hu, Hsing-Kuo Hua, Hongzhan Huang, Robel Kahsay, Raja Mazumder, Peter McGarvey, Darren Natale, Anastasia Nikolskaya, Natalia Petrova, Baris Suzek, Sona Vasudevan, C. R. Vinayaka, Lai Su Yeh, Xin Yuan and Jian Zhang at the Protein Information Resource (PIR)

The authors wish it to be know that, in their opinion, all authors should be regarded as joint First Authors different uses. The UniProt Knowledgebase (UniProtKB) is an expertly curated database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) is a comprehensive sequence repository, reflecting the history of all protein sequences (1). UniProt Reference Clusters (UniRef) merge closely related sequences based on sequence identity to speed up searches. UniProt is built upon the extensive bioinformatics infrastructure and scientific expertise at European Bioinformatics Institute (EBI), Protein Information Resource (PIR) and Swiss Institute of Bioinformatics (SIB). It is freely and easily accessible by researchers to conduct interactive and custom-tailored analyses for proteins of interest to facilitate hypothesis generation and knowledge discovery.

CONTENT

UniProtKB

UniProtKB consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. The former contains manually annotated records with information extracted from literature and curator-evaluated computational analysis. To achieve accuracy, annotations are performed by biologists with specific expertise. Information including function, catalytic activity, subcellular location, disease, structure and post-translational modifications is annotated. An important part of the annotation process involves the merging of different reports for a single protein. After a careful inspection of the sequences, the annotator selects the reference sequence, does the corresponding merging and lists the splice and genetic variants along with disease information when available. Any discrepancies between the different sequence sources are also annotated. Cross-references are provided to the underlying nucleotide sequence sources as well as to many other useful databases including organism-specific, domain, family and disease databases. UniProtKB/TrEMBL contains high quality computationally analyzed records enriched with automatic annotation and classification. The computer-assisted annotation is created using automatically generated rules as in Spearmint (2) or manually curated rules based on protein families, including HAMAP family rules (3), RuleBase rules (4) and PIRSF classification-based name rules and site rules (5,6). UniProtKB/TrEMBL contains the translations of all coding sequences present in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases, the sequences of PDB structures and data derived from amino acid sequences that are directly submitted to the UniProtKB or scanned from the literature. We exclude some types of data such as DDBJ/EMBL/DDBJ entries that encode small fragments, synthetic sequences, most non-germline immunoglobulins and T-cell receptors, most patent sequences and some highly over-represented data. Records are selected for full manual annotation and integration into UniProtKB/Swiss-Prot according to defined annotation priorities.

UniRef

The UniRef databases provide three clustered sets (UniRef100, 90 and 50) of sequences from UniProtKB and selected UniParc records in order to obtain complete coverage of sequence space at several resolutions while hiding redundant sequences from view. The UniRef100 database combines identical sequences and sub-fragments with 11 or more residues into a single UniRef entry, which displays the sequence of a representative protein, with the accession numbers of all the UniProtKB entries within the cluster and links to the corresponding UniProtKB and UniParc records. UniRef90 and UniRef50 are built by further clustering UniRef100 sequences with 11 or more residues using the CD-HIT algorithm (7) such that each cluster is composed of sequences that have at least 90 or 50% sequence identity, respectively, to the representative sequence. Selection of the representative sequence in each UniRef cluster is based on the ranking of all the sequences in the cluster using the following criteria in descending precedence:

(i) Quality of the entry: member entries from UniProtKB/Swiss-Prot section are preferred.
(ii) Meaningful name: entries with names that do not contain non-biological or non-descriptive words, such as hypothetical, probable, are preferred.
(iii) Organism: entries from model organisms are preferred.
(iv) Length of the sequence: longest sequence are preferred.

The UniRef databases are generated based on the UniProtKB and UniParc databases, thus providing up-to-date collections of sequences. UniRef100 is the most comprehensive and non-redundant protein sequence dataset. UniRef90 and UniRef50 yield a database size reduction of ∼40 and 65%, respectively, providing for significantly faster sequence similarity searches. In addition, UniRef databases reduce the bias in sequence searches by providing a more even sampling of sequence space.

UniParc

UniParc is the main sequence storehouse and is a comprehensive repository that reflects the history of all protein sequences (1). UniParc houses all new and revised protein sequences from various sources to ensure that complete coverage is available at a single site. It includes not only UniProtKB but also translations from the EMBL-Bank/DDBJ/GenBank Nucleotide Sequence Databases, the Ensembl database of animal genomes, the International Protein Index (IPI), the Protein Data Bank (PDB), NCBI's Reference Sequence Collection (RefSeq), model organism databases FlyBase and WormBase and protein sequences from the European, American and Japanese Patent Offices. To avoid redundancy, sequences are handled as strings—all sequences 100% identical over the entire length are merged, regardless of source organism. New and updated sequences are loaded on a daily basis, cross-referenced to the source database accession number and provided with a sequence version that increments upon changes to the underlying sequence. The basic information stored within each UniParc entry is the identifier, the sequence, cyclic redundancy check number, source database(s) with accession and version numbers and a time stamp. In addition, each source database accession number is tagged with its status in that database, indicating if the sequence still exists or has been deleted in the source database. UniParc records are designed to be without annotation since the annotation will be only true in the real biological context of the sequence: proteins with the same sequence may have different functions depending on species, tissue, developmental stage, etc.

The UniRef databases are generated based on the UniProtKB and UniParc databases, thus providing up-to-date collections of sequences. UniRef100 is the most comprehensive and non-redundant protein sequence dataset. UniRef90 and UniRef50 yield a database size reduction of ∼40 and 65%, respectively, providing for significantly faster sequence similarity searches. In addition, UniRef databases reduce the bias in sequence searches by providing a more even sampling of sequence space.

NEW FEATURES

UniSave

The UniProtKB Sequence/Annotation Version database (UniSave) is a comprehensive archive of UniProtKB entry versions (8). All changed UniProtKB/Swiss-Prot and UniProt/TrEMBL entries are added to UniSave on a bi-weekly basis to coincide with the UniProtKB releases. UniSave is available at http://www.ebi.ac.uk/uniprot/unisave.

ID mapping

UniProt provides a mapping service to convert common gene IDs and protein IDs to UniProtKB AC/ID and vice versa. Mappings are either inherited from cross-references within UniProtKB entries or are based on the existing mappings between EMBL and GenBank entries, while others make use of cross-references obtained from the iProClass database (9). This service is available at http://www.uniprot.org/search/idmapping.shtml, where users can map between UniProtKB and >30 other data sources such as NCBI (e.g. gi numbers, RefSeq accession numbers, Entrez Gene IDs, PubMed IDs), GO (www.geneontology.org/), PFAM (www.sanger.ac.uk/Software/Pfam/) and PIRSF (pir.georgetown.edu/pirsf.shtml). In addition, users can also download selected mappings in the form of a tab-delimited table from ftp://ftp.pir.georgetown.edu/databases/iproclass/.

Format changes

Recent format changes. A number of UniProtKB format changes have recently been introduced to improve data consistency:

(i) The DT line (DaTe) changed from showing only the dates corresponding to full UniProtKB releases to displaying the date of the bi-weekly release at which an entry is integrated or updated. The information concerning the release number has been dropped and the entry and sequence version numbers in the DT lines were introduced instead. The sequence version number of an entry is incremented by one when its amino acid sequence is modified, whereas the entry version number is incremented by one whenever any data in the flat file representation of the entry is modified.

(ii) A new line type has been introduced to viral entries to indicate the host(s) either as a specific organism or taxonomic group of organisms. This line has been termed OH for Organism Host and contains the host name and taxonomy ID.

(iii) The CC line (Comment) topic DATABASE has been replaced by WEB RESOURCE to clarify the conceptual difference between the content of these lines and the DR (Database cross-Reference) lines.

(iv) Pre-translational events have so far been represented by several feature keys. To improve the consistency of annotation of pre- and co-translational events, the feature key VARSPLIC was removed and the new feature key VAR_SEQ created for the description of alternative splicing, alternative promoter usage, alternative initiation and ribosomal frameshifting.

Forthcoming format changes.

(i) The format of the ID line will be changed to better reflect the annotation status of an entry. The current STANDARD and PRELIMINARY data classes will be replaced by 'Reviewed' (entries that have been manually reviewed and annotated by UniProtKB curators) and 'Unreviewed' (computer-annotated entries that have not been reviewed by UniProtKB curators), respectively. In addition, the MoleculeType field, which is a legacy of compatibility with the EMBL flat file format, will be dropped.

(ii) Since in most cases protein sequences are derived from translation of nucleotide sequences and there may or may not be definitive experimental evidence for their existence, a new line type will be introduced to indicate the evidence for the existence of a protein (PE line). The PE line will have one of the following values: 'evidence at protein level', 'evidence at transcript level', 'inferred from homology' or 'predicted'. Unreviewed entries will have an additional value, 'unassigned'. However, it should be noted that the PE line will not give information on the correctness of the sequence.

(iii) As mentioned before, the feature key INIT_MET is only used to indicate that the initiator methionine has been cleaved off. Currently, the initiator methionine is not included in the sequence of a UniProtKB entry in such a case and the INIT_MET sequence coordinates are therefore 0. The initiator methionine will be added back to such protein sequences and the sequence coordinates of the feature key INIT_MET accordingly changed to 1.

(iv) The FASTA header line of UniProtKB and UniRef entries will be standardized. In the former case the format will consist of >UniqueIdentifier|EntryName ProteinName—OrganismName, whereas in the latter it will follow the following format: >UniqueIdentifier Cluster: ClusterName; n = Members; Taxon|Rep: ProteinName—OrganismName.

- For UniProtKB, the UniqueIdentifier is the primary accession number of the UniProtKB entry or, in the case of entries that describe several protein isoforms, an isoform identifier; EntryName is the entry name of the UniProtKB entry; ProteinName is the recommended or submitted protein name of the UniProtKB entry (this is the name before the first bracket, excluding 'precursor' but including 'Fragment' if

appropriate); OrganismName is the scientific name of the organism of the UniProtKB entry. Examples:
>P24856|ANP_NOTCO Ice-structuring glycoprotein (Fragment)—*Notothenia coriiceps neglecta*
>P51650-1|SSDH_RAT Succinate semialdehyde dehydrogenase—*Rattus norvegicus*

- For UniRef, the UniqueIdentifier is the primary accession number of the UniRef cluster; ClusterName is the name of the UniRef cluster; Members is the number of UniRef cluster members; Taxon is the scientific name of the lowest common taxon shared by all UniRef cluster members; ProteinName is the protein name of the representative member of the UniRef cluster; OrganismName is the scientific name of the organism of the representative member of the UniRef cluster. Example:
  >UniRef50_P24856 Cluster: Ice-structuring glycoprotein (Fragment); $n = 15$; Holacanthopterygii|Rep: Ice-structuring glycoprotein (Fragment)—*Notothenia coriiceps neglecta*
- UniParc is not represented here as its header is purely the UniParc accession number.

RECENT CHANGES

New documents

A number of documents, available both by ftp and on the Web site, have been added. The document nameprot.txt lists a number of rules for naming proteins to standardize the nomenclature for a given protein across related organisms, with the hope that authors/laboratories will follow as much as possible these rules for naming new proteins. The document orysa.txt lists all the Oryza sativa (rice) entries of the UniProtKB/Swiss-Prot section with the corresponding chromosome locus, the UniProtKB accession number, entry name, the description and the gene name(s). The document scorpktx.txt lists the potassium-channel-specific scorpion toxins known to date (10,11) along with the UniProtKB accession number, the entry name and the systematic name. Finally, the document ptmlist.txt contains the controlled vocabulary and associated feature keys for post-translational modifications.

Metagenomic and environmental sequences

Swiss-Prot and TrEMBL sections of the UniProtKB contain entries with a known taxonomic source. However, a new development in sequence production—namely, the availability of metagenomic data—has necessitated the creation of a separate section, UniProt Metagenomic and Environmental Sequences (UniMES).

UPCOMING DEVELOPMENTS

Annotation of UniProtKB/TrEMBL entries

To improve the quality of the protein names in UniProtKB/TrEMBL, an effort is underway to start manually curating protein names using protein PIRSFs and their accompanying name rules and/or site rules (5,6). It is important to note that all the protein names will be checked manually prior to updating the individual UniProtKB/TrEMBL record. In the context of the Human Proteome Initiative annotation program, the protein names and gene symbols of representative human entries in UniProtKB/TrEMBL will be updated so as to provide the community with a cleaner human proteome set that encompasses the 15 000 reviewed human entries in UniProtKB/Swiss-Prot and about 5000 as yet unreviewed human UniProtKB/TrEMBL entries.

DATABASE ACCESS AND FEEDBACK

UniProt is freely available for both commercial and non-commercial use. Please see http://www.uniprot.org/terms for details. The UniProt databases can be accessed online (http://www.uniprot.org) or downloaded in several formats (ftp://ftp.uniprot.org/pub). New releases are published on a bi-weekly basis.

ACKNOWLEDGEMENTS

UniProt is mainly supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01. Additional support for the EBI's involvement in UniProt comes from the European Commission (EC)'s FELICS grant (021902RII3) and from the NIH grant 1R01HGO2273-01. UniProtKB/Swiss-Prot activities at the SIB are supported by the Swiss Federal Government through the Federal Office of Education and Science. PIR activities are also supported by the NIH grants for NIAID proteomic resource (HHSN266200400061C) and grid enablement (NCI-caBIGICR-10-10-01) and BioThesaurus (ITR-0205470). Funding to pay the Open Access publication charges for this article was provided by the UniProt NIH grant 1 U01 HG02712-01.

Conflict of interest statement. None declared.

REFERENCES

1. Leinonen,R., Diez,F.G., Binns,D., Fleischmann,W., Lopez,R. and Apweiler,R. (2004) UniProt archive. *Bioinformatics*, **20**, 3236–3237.
2. Wieser,D., Kretschmann,E. and Apweiler,R. (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20**, i342–i347.
3. Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.
4. Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
5. Wu,C.H., Nikolskaya,A., Huang,H., Yeh,L.S., Natale,D.A., Vinayaka,C.R., Hu,Z.Z., Mazumder,R., Kumar,S., Kourtesis,P. *et al.* (2004) PIRSF: family classification system at the protein information resource. *Nucleic Acids Res.*, **32**, D112–D114.
6. Natale,D.A., Vinayaka,C.R. and Wu,C.H. (2004) Large-scale, classification-driven, rule-based functional annotation of proteins. In Subramaniam,S. (ed.), *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. Bioinformatics Volume*, John Wiley & Sons, Ltd.
7. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
8. Leinonen,R., Nardone,F., Zhu,W. and Apweiler,R. (2006) UniSave: the UniProtKB sequence/annotation version database. *Bioinformatics*, **22**, 1284–1285.

9. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z. and Barker,W.C. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.
10. Tytgat,J., Chandy,K.G., Garcia,M.L., Gutman,G.A., Martin-Eauclaire,M.F., van der Walt,J.J. and Possani,L.D. (1999) A unified nomenclature for short-chain peptides isolated from scorpion venoms: alpha-KTx molecular subfamilies. *Trends Pharmacol. Sci.*, **20**, 444–447.
11. Rodriguez de la Vega,R.C. and Possani,L.D. (2004) Current views on scorpion toxins specific for K+-channels. *Toxicon*, **43**, 865–875.